Retrieval-Enhanced Machine Learning: Synthesis and Opportunities

Fernando Diaz
Carnegie Mellon University
Pittsburgh, PA, United States
diazf@acm.org

Andrew Drozdov*
University of Massachusetts
Amherst, MA, United States
adrozdov@cs.umass.edu

To Eun Kim
Carnegie Mellon University
Pittsburgh, PA, United States
toeunk@cs.cmu.edu

Alireza Salemi
University of Massachusetts
Amherst, MA, United States
asalemi@cs.umass.edu

Hamed Zamani
University of Massachusetts
Amherst, MA, United States
zamani@cs.umass.edu

Abstract

Retrieval-enhanced machine learning (REML) refers to the use of information retrieval methods to support reasoning and inference in machine learning tasks. Although relatively recent, these approaches can substantially improve model performance. This includes improved generalization, knowledge grounding, scalability, freshness, attribution, interpretability and on-device learning. To date, despite being influenced by work in the information retrieval community, REML research has predominantly been presented in natural language processing (NLP) conferences. Our tutorial addresses this disconnect by introducing core REML concepts and synthesizing the literature from various domains in machine learning (ML), including, but beyond NLP. What is unique to our approach is that we used consistent notations, to provide researchers with a unified and expandable framework. The tutorial will be presented in lecture format based on an existing manuscript, with supporting materials and a comprehensive reading list available at https://retrieval-enhanced-ml.github.io/sigir-ap2024-tutorial.

CCS Concepts

• Information systems → Information retrieval; **• Computing methodologies → Machine learning**.

Keywords

Information Retrieval, Machine Learning

ACM Reference Format:
Fernando Diaz, Andrew Drozdov, To Eun Kim, Alireza Salemi, and Hamed Zamani. 2024. Retrieval-Enhanced Machine Learning: Synthesis and Opportunities. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region (SIGIR-AP '24), December 9–12, 2024, Tokyo, Japan.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3673791.3698439

*Now at Databricks.

SIGIR-AP '24, December 9–12, 2024, Tokyo, Japan

ACM ISBN 979-8-4007-0724-7/24/12
https://doi.org/10.1145/3673791.3698439

1 Motivation

Retrieval systems, originally designed for human use, are increasingly being integrated into machine learning models to extend their access to information beyond fixed model parameters [21]. These systems can act as an external 'memory', using mechanisms like nearest neighbor databases or keyword queries. Recent empirical evidence shows that incorporating retrieval systems enhances model performance, improving generalization, knowledge grounding, scalability, freshness, attribution, and on-device learning [39].

In light of the success of these methods, Zamani et al. [39] introduced retrieval-enhanced machine learning (REML), a research program focused on the development of information retrieval techniques for artificial intelligence systems. In the year since it was published, the machine learning and natural language processing communities have continued to make progress in the design of REML [2, 4, 7, 28].

While effective, much of the current REML research has been disconnected from the Information Retrieval (IR) community. As a result, many of the insights from existing IR research remains underutilized. For example, when retrieval methods are used, they are often simple approaches such as BM25. At the same time, retrieval methods to date have–with some exception–focused on their use by people as end users, not models.

This tutorial explores the integration of retrieval systems into machine learning models. We will cover the historical and contemporary use of retrieval in machine learning and synthesizes methods across domains using consistent mathematical notation, which is lacking in current literature. Instead of organizing the tutorial by applications, the tutorial is structured by the components of the REML framework: *Querying*, *Searching*, *Presentation*, *Consumption*, *Storing*, *Optimization*, and *Evaluation*. This structure enhances understanding of each component's role and interaction within the framework. Additionally, this allows us to highlight core functionalities and generalize to new domains.

To this end, we will present examples of various tasks—both within and beyond natural language processing—at different levels of granularity where REML is applied. Additionally, we will show

Authors are listed in alphabetical order.

Time (min)	Topic	In Manuscript
20	Introduction	Sec. 1-2
20	Querying	Sec. 3
10	Searching	Sec. 4
30	Presentation & Consumption	Sec. 5
10	Q & A	
30	Storing	Sec. 6
20	Optimization	Sec. 7
15	Evaluation	Sec. 8
15	Future Direction & Conclusion	Sec. 9-10
10	Q & A	
3 hours		

Table 1: Tutorial schedule and corresponding manuscript sections [21].

how different types of knowledge aid in generalization and address the computational costs associated with REML.

2 Objectives

The goal of this tutorial is to provide information retrieval researchers with a clear, formal description of the various REML approaches so that they can quickly begin research in the area. As such, this tutorial will have the following objectives, (i) survey and synthesize the variety of REML approaches based on common strategies, (ii) connect abstract themes to existing information retrieval research, and (iii) outline a set of new open research problems for the information retrieval and ML community. This tutorial will formally define the various strategies for retrieval enhancement using consistent notation, allowing researchers easy entry to the field.

3 Relevance to the Community

The information retrieval community has historically collaborated with peers in natural language processing to support tasks like question-answering [36]. Research to better support these tasks is regularly published at SIGIR and studied at fora like TREC.

As described by Zamani et al. [39], REML can be seen as an update and generalization of this thread of IR research. It updates classic domains like retrieval-based question-answering by integrating them into modern deep learning architectures. It generalizes across these domains by recognizing that retrieval does not have to be constrained to be a single stage in reasoning (e.g. as a candidate generation step). Moreover, retrieval does not have to be constrained to text corpora and can support knowledge and memory of abstract representations and concepts. Our tutorial provides a theoretical unification across common themes in REML, and highlights opportunities for future research.

Drawing on our previous experience of hosting a workshop on REML at SIGIR 2023 [4], where we witnessed substantial interest, we have integrated insights from those discussions into this tutorial. The growing popularity of REML is also evident in the recent surge of research papers, open-source projects, and industry applications that employ this innovative approach. As such, a comprehensive tutorial on REML is both timely and valuable for the IR community.

4 Detailed Schedule

Table 1 presents the overall schedule along with the corresponding sections from the manuscript [21]. The detailed content for each section, along with representative papers, is provided in the bullet points below.

1. **Introduction** (presenter: Diaz)
   - Prehistory and definition of REML.
   - Motivations: generalization, knowledge grounding, scalability, freshness, attribution, and on-device learning [12, 17, 22, 39].
   - Applications beyond NLP (e.g. image generation [6], image classification [23], protein structure prediction [16])
2. **Querying** (presenter: Salemi)
   - *How query spaces are represented and constructed by predictive models.*
   - Input Reformulation: compression [18, 27], expansion [37, 43], and conversion [34].
   - Input Decomposition [25, 42].
   - Unified equation for Querying.
3. **Searching** (presenter: Salemi)
   - *How queries and stored items are combined to construct retrieval results.*
   - Sparse [10], dense [19], and reranking [26] models.
   - Generative retrievers [35, 40].
   - Unified equation for Searching.
4. **Presentation & Consumption** (presenter: Drozdov)
   - *How retrieval results are represented and consumed by the predictive models.*
   - Presentation: transformation [11], composition [33], and truncation [13].
   - Consumption with different granularities [15], algorithms [3], efficiency [8], and attribution [11].
   - Unified equation for Presentation and Consumption.
5. **Storing** (presenter: Kim)
   - *How retrievable items are represented and indexed.*
   - Storage operations (construction and management).
   - Coupled [12, 24] and Decoupled [1, 17] storage.
   - Unified equation for Storing.
6. **Optimization** (presenter: Zamani)
   - *How retrieval models use feedback provided by predictive models to update their parameters, and how predictive models are optimized for performance.*
   - Conditional optimization of retrieval [14, 38] or predictive models [5].
   - Joint optimization of retrieval and predictive models [13, 22, 41].
7. **Evaluation** (presenter: Diaz)
   - *How REML components are benchmarked.*
   - Extrinsic and Intrinsic evaluations [9, 29].
8. **Future Direction & Conclusion** (presenter: Diaz)
   - Future directions of each component of REML [20, 30–32].

5 Supporting Materials

This tutorial builds on material and structure from an existing manuscript written by the organizers [21]. Attendees will receive slides, including an annotated bibliography, throughout the session.

Detailed information and a comprehensive reading list can be found at the following link: https://retrieval-enhanced-ml.github.io/sigir-ap2024-tutorial/.

6 Presenters

6.1 Fernando Diaz

Fernando Diaz is an Associate Professor at Carnegie Mellon's Language Technologies Institute (LTI), focusing on the design of information access systems such as search engines, music recommendation services, and crisis response platforms. His research also explores the societal implications of artificial intelligence. Previously, he was the assistant managing director of Microsoft Research Montréal, leading the FATE team, and a director of research at Spotify. His work has been recognized with awards from SIGIR, CIKM, CSCW, WSDM, ISCRAM, and ECIR. Fernando is a recipient of the 2017 British Computer Society Karen Spärck Jones Award and holds a CIFAR AI Chair. He has co-organized NIST TREC tracks, WSDM (2013), Strategic Workshop on Information Retrieval (2018), FAT* (2019), SIGIR (2021), and the CIFAR Workshop on Artificial Intelligence and the Curation of Culture (2019). He earned his BS from the University of Michigan and his MS and PhD from the University of Massachusetts Amherst.

6.2 Andrew Drozdov

Andrew Drozdov is a research scientist at Databricks, specializing in building advanced systems for retrieval-augmented generation (RAG). He serves as an Area Chair for Information Retrieval and Efficient NLP tracks at ACL Rolling Review. His previous roles include research internships at Google Research's Brain Team and IBM Research's Multilingual NLP Group. Andrew earned his PhD from the University of Massachusetts Amherst, co-advised by Professors Andrew McCallum and Mohit Iyyer, and holds a BS from the University of Michigan and an MS from New York University, where he collaborated with Professors Sam Bowman and Kyunghyun Cho.

6.3 To Eun Kim

To Eun Kim is a PhD student at the Language Technologies Institute (LTI) at Carnegie Mellon University, where he is advised by Professor Fernando Diaz. His research focuses on retrieval-enhanced machine learning, with a recent emphasis on algorithmic fairness in REML models and improving known-item retrieval with large language models. He holds an MEng in Computer Science from University College London (UCL), where he worked with Professor Emine Yilmaz and Professor Aldo Lipani, and was a lead author in the Alexa Prize TaskBot Challenge.

6.4 Alireza Salemi

Alireza Salemi is a PhD student at the University of Massachusetts Amherst, where he is advised by Professor Hamed Zamani and works as a research assistant at the Center for Intelligent Information Retrieval (CIIR). His research focuses on both uni- and multi-modal retrieval-enhanced machine learning. He also works on multi-modal knowledge-intensive visual question answering, personalizing pre-trained language models, and developing retrieval-enhanced architectures. Alireza has authored several papers in the domain of retrieval-enhanced machine learning, including at SIGIR 2024. He holds a BS in Computer Engineering from the University of Tehran.

6.5 Hamed Zamani

Hamed Zamani is an Associate Professor at the University of Massachusetts Amherst, where he also serves as the Associate Director of the Center for Intelligent Information Retrieval (CIIR). Prior to UMass, he was a Researcher at Microsoft working on search and recommendation problems. His research focuses on designing and evaluating (interactive) information access systems, including search engines, recommender systems, and question answering. His work has led to over 90 refereed publications in the field, including some recent work on the topic of REML. His research has received a few Best Paper and Honorable Mentions from SIGIR, CIKM, and ICTIR. He is a recipient of the NSF CAREER Award and Amazon Research Award. He is an Associate Editor of the ACM Transactions on Information Systems (TOIS), organized multiple workshops at SIGIR, RecSys, WSDM, WWW, and KDD conferences, and presented multiple tutorials at SIGIR and WWW.

Acknowledgments

This work was supported in part by the Center for Intelligent Information Retrieval, in part by NSF grant numbers 2402873 and 2402874, and in part by the Office of Naval Research contract number N000142412612. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsors.

References

[1] Uri Alon, Frank F. Xu, Junxian He, Sudipta Sengupta, Dan Roth, and Graham Neubig. 2022. Neuro-Symbolic Language Modeling with Automaton-augmented Retrieval. In *ICML 2022 Workshop on Knowledge Retrieval and Language Models*. https://openreview.net/forum?id=ZJZmKGM6UB

[2] Akari Asai, Sewon Min, Zexuan Zhong, and Danqi Chen. 2023. ACL 2023 Tutorial: Retrieval-based Language Models and Applications. *ACL 2023* (2023).

[3] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=hSyW5go0v8

[4] Michael Bendersky, Danqi Chen, Fernando Diaz, and Hamed Zamani. 2023. SIGIR 2023 Workshop on Retrieval Enhanced Machine Learning (REML @ SIGIR 2023). In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23)*. Association for Computing Machinery, 3468–3471. https://doi.org/10.1145/3539618.3591925

[5] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to Answer Open-Domain Questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 1870–1879. https://doi.org/10.18653/v1/P17-1171

[6] Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W. Cohen. 2022. Re-Imagen: Retrieval-Augmented Text-to-Image Generator. *ArXiv* abs/2209.14491 (2022).

[7] Rajarshi Das, Patrick Lewis, Sewon Min, June Thai, and Manzil Zaheer (Eds.). 2022. *Proceedings of the 1st Workshop on Semiparametric Methods in NLP: Decoupling Logic from Knowledge*. Association for Computational Linguistics. https://aclanthology.org/2022.spanlp-1.0

[8] Michiel De Jong, Yury Zemlyanskiy, Nicholas FitzGerald, Joshua Ainslie, Sumit Sanghai, Fei Sha, and William W. Cohen. 2023. Pre-Computed Memory or on-the-Fly Encoding? A Hybrid Approach to Retrieval Augmentation Makes the Most of Your Compute. In *Proceedings of the 40th International Conference on Machine Learning (ICML'23)*. JMLR.org, Article 290, 14 pages.

[9] Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. RAGAs: Automated Evaluation of Retrieval Augmented Generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational*

Linguistics: System Demonstrations. Association for Computational Linguistics, 150–158. https://aclanthology.org/2024.eacl-demo.16

[10] Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. SPLADE: Sparse Lexical and Expansion Model for First Stage Ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21)*. Association for Computing Machinery, 2288–2292. https://doi.org/10.1145/3404835.3463098

[11] Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling Large Language Models to Generate Text with Citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 6465–6488. https://doi.org/10.18653/v1/2023.emnlp-main.398

[12] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: Retrieval-Augmented Language Model Pre-Training. In *Proceedings of the 37th International Conference on Machine Learning (ICML'20)*. JMLR.org, Article 368, 10 pages.

[13] Sebastian Hofstätter, Jiecao Chen, Karthik Raman, and Hamed Zamani. 2023. Fid-light: Efficient and effective retrieval-augmented text generation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1437–1447.

[14] Gautier Izacard and Edouard Grave. 2021. Distilling Knowledge from Reader to Retriever for Question Answering. In *International Conference on Learning Representations*. https://openreview.net/forum?id=NTEz-6wysdb

[15] Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active Retrieval Augmented Generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 7969–7992. https://doi.org/10.18653/v1/2023.emnlp-main.495

[16] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. 2021. Highly accurate protein structure prediction with AlphaFold. *nature* 596, 7873 (2021), 583–589.

[17] Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Generalization through Memorization: Nearest Neighbor Language Models. In *International Conference on Learning Representations*. https://openreview.net/forum?id=HklBjCEKvH

[18] Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Dan Roth. 2017. Learning What is Essential in Questions. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*. Association for Computational Linguistics, 80–89. https://doi.org/10.18653/v1/K17-1010

[19] Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20)*. Association for Computing Machinery, 39–48. https://doi.org/10.1145/3397271.3401075

[20] To Eun Kim and Fernando Diaz. 2024. Towards Fair RAG: On the Impact of Fair Ranking in Retrieval-Augmented Generation. arXiv:2409.11598 [cs.IR] https://arxiv.org/abs/2409.11598

[21] To Eun Kim, Alireza Salemi, Andrew Drozdov, Fernando Diaz, and Hamed Zamani. 2024. Retrieval-Enhanced Machine Learning: Synthesis and Opportunities. arXiv:2407.12982 [cs.LG] https://arxiv.org/abs/2407.12982

[22] Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

[23] Alexander Long, Wei Yin, Thalaiyasingam Ajanthan, Vu Nguyen, Pulak Purkait, Ravi Garg, Alan Blair, Chunhua Shen, and Anton van den Hengel. 2022. Retrieval Augmented Classification for Long-Tail Visual Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 6959–6969.

[24] Bodhisattwa Prasad Majumder, Bhavana Dalvi, Peter Jansen, Oyvind Tafjord, Niket Tandon, Li Zhang, Chris Callison-Burch, and Peter Clark. 2024. CLIN: A Continually Learning Language Agent for Rapid Task Adaptation and Generalization. https://openreview.net/forum?id=d5DGVHMdsC

[25] Sewon Min, Victor Zhong, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2019. Multi-hop Reading Comprehension through Question Decomposition and Rescoring. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 6097–6109. https://doi.org/10.18653/v1/P19-1613

[26] Bhaskar Mitra and Nick Craswell. 2018. An Introduction to Neural Information Retrieval. *Found. Trends Inf. Retr.* 13, 1 (dec 2018), 1–126. https://doi.org/10.1561/1500000061

[27] Ryan Musa, Xiaoyan Wang, Achille Fokoue, Nicholas Mattei, Maria Chang, Pavan Kapanipathi, Bassem Makni, Kartik Talamadupula, and Michael Witbrock. 2019. Answering Science Exam Questions Using Query Reformulation with Background Knowledge. In *Automated Knowledge Base Construction (AKBC)*. https://openreview.net/forum?id=HJxYZ-5paX

[28] Maithra Raghu, Urvashi Khandelwal, Chiyuan Zhang, Matei Zaharia, and Alexander Rush (Eds.). 2022. *Workshop on Knowledge Retrieval and Language Models*. ICML.

[29] Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2024. ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Association for Computational Linguistics, 338–354.

[30] Alireza Salemi, Surya Kallumadi, and Hamed Zamani. 2024. Optimization Methods for Personalizing Large Language Models through Retrieval Augmentation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24)*. Association for Computing Machinery, 752–762. https://doi.org/10.1145/3626772.3657783

[31] Alireza Salemi and Hamed Zamani. 2024. Evaluating Retrieval Quality in Retrieval-Augmented Generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24)*. Association for Computing Machinery, 2395–2400. https://doi.org/10.1145/3626772.3657957

[32] Alireza Salemi and Hamed Zamani. 2024. Towards a Search Engine for Machines: Unified Ranking for Multiple Retrieval-Augmented Large Language Models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24)*. Association for Computing Machinery, 741–751. https://doi.org/10.1145/3626772.3657733

[33] Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D Manning. 2024. RAPTOR: Recursive Abstractive Processing for Tree-Organized Retrieval. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=GN921JHCRw

[34] Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language Models Can Teach Themselves to Use Tools. In *Thirty-seventh Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=Yacmpz84TH

[35] Yi Tay, Vinh Tran, Mostafa Dehghani, Jianmo Ni, Dara Bahri, Harsh Mehta, Zhen Qin, Kai Hui, Zhe Zhao, Jai Gupta, Tal Schuster, William W Cohen, and Donald Metzler. 2022. Transformer Memory as a Differentiable Search Index. In *Advances in Neural Information Processing Systems*, Vol. 35. Curran Associates, Inc., 21831–21843. https://proceedings.neurips.cc/paper_files/paper/2022/file/892840a6123b5ec99ebaab8be1530fba-Paper-Conference.pdf

[36] Ellen M. Voorhees. 2003. Overview of the TREC 2003 Question Answering Track. In *Proceedings of the Twelfth Text REtrieval Conference*.

[37] Liang Wang, Nan Yang, and Furu Wei. 2023. Query2doc: Query Expansion with Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 9414–9423. https://doi.org/10.18653/v1/2023.emnlp-main.585

[38] Sohee Yang and Minjoon Seo. 2020. Is Retriever Merely an Approximator of Reader? *CoRR* abs/2010.10999 (2020). arXiv:2010.10999 https://arxiv.org/abs/2010.10999

[39] Hamed Zamani, Fernando Diaz, Mostafa Dehghani, Donald Metzler, and Michael Bendersky. 2022. Retrieval-Enhanced Machine Learning. In *Proceedings of the 45th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[40] Hansi Zeng, Chen Luo, Bowen Jin, Sheikh Sarwar, Tianxin Wei, and Hamed Zamani. 2024. Scalable and Effective Generative Information Retrieval. In *Proceedings of the 2024 ACM Web Conference (WWW '24)*. Association for Computing Machinery.

[41] Yizhe Zhang, Siqi Sun, Xiang Gao, Yuwei Fang, Chris Brockett, Michel Galley, Jianfeng Gao, and Bill Dolan. 2022. Retgen: A joint framework for retrieval and grounded text generation modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 11739–11747.

[42] Ben Zhou, Kyle Richardson, Xiaodong Yu, and Dan Roth. 2022. Learning to Decompose: Hypothetical Question Decomposition Based on Comparable Texts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2223–2235. https://doi.org/10.18653/v1/2022.emnlp-main.142

[43] Yunchang Zhu, Liang Pang, Yanyan Lan, Huawei Shen, and Xueqi Cheng. 2021. Adaptive Information Seeking for Open-Domain Question Answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 3615–3626. https://doi.org/10.18653/v1/2021.emnlp-main.293